WHAT IS CLAIMED IS:

1. A method for detecting similar objects in a collection of such objects, comprising, for each of two objects:

modifying a previous method for detecting similar objects so that memory requirements are reduced while avoiding false detections approximately as well as in the previous method, wherein the modifying comprises:

combining a number of samples of features into each of a total number of supersamples, wherein the number of samples is reduced from a number of samples used in the previous method;

recording each of the total number of supersamples to a number of bits of precision, wherein the number of bits of precision is reduced from a number of bits of precision used in the previous method; and

requiring a number of matching supersamples out of the total number of supersamples in order to conclude that the two objects are sufficiently similar, wherein the number of matching supersamples is greater than a number of matching supersamples required in the previous method.

2. The method of claim 1 wherein requiring the number of matching supersamples comprises requiring all but one of the total number of supersamples to match.

3. The method of claim 1 wherein requiring the number of matching supersamples comprises requiring all but two of the total number of supersamples to match.

4. The method of claim 1 wherein requiring the number of matching supersamples comprises requiring all supersamples to match.

5. The method of claim 1 wherein combining the number of samples into each of the total number of supersamples comprises combining four samples into each of the total number of supersamples, wherein the number of samples used in the previous method is 14.

6. The method of claim 5 wherein:

recording each supersample to the first number of bits of precision comprises recording each supersample to 16 bits of precision, wherein the second number of bits of precision used in the previous method is 64; and

requiring the number of matching supersamples comprises requiring four supersamples of six to match, wherein the number of matching supersamples required in the previous method is two supersamples of six.

7. The method of claim 5 wherein requiring the number of matching supersamples comprises requiring five supersamples of seven to match, wherein the number of matching supersamples required in the previous method is two supersamples of six.

8. The method of claim 1 wherein the objects are documents, and the method is used in association with a search engine query service to determine clusters of query results that are near-duplicate documents.

9. The method of claim 8, further comprising selecting a single document in each cluster to report.

10. The method of claim 9 wherein selecting the single document is by way of a ranking function.

11. A method for determining groups of near-duplicate items in a search engine query result, comprising, for each of two items being compared:

combining four samples of features into each of six supersamples;

recording each supersample to 16 bits of precision; and

requiring four of the six supersamples to match.

12. The method of claim 11, further comprising selecting a single document in each cluster to report.

13. The method of claim 12 wherein selecting the single document is by way of a ranking function.

14. A method for determining groups of near-duplicate items in a search engine query result, comprising, for each of two items being compared:

combining four samples of features into each of seven supersamples;

recording each supersample to 16 bits of precision; and

requiring five of the seven supersamples to match.

15. The method of claim 14, further comprising selecting a single document in each cluster to report.

16. The method of claim 15 wherein selecting the single document is by way of a ranking function.

17. A computer-readable medium embodying machine instructions implementing a current method for detecting similar objects in a collection of such objects, wherein the current method comprises modification of a previous method for detecting similar objects so that memory requirements are reduced while avoiding false detections approximately as well as in the previous method, the current method comprising:

combining a number of samples of features into each of a total number of supersamples, wherein the number of samples is reduced from a number of samples used in the previous method;

recording each of the total number of supersamples to a number of bits of precision, wherein the number of bits of precision is reduced from a number of bits of precision used in the previous method; and

requiring a number of matching supersamples in order to conclude that the two objects are sufficiently similar, wherein the number of matching supersamples is greater than a number of matching supersamples required in the previous method.

18. The computer-readable medium of claim 17 wherein requiring the number of matching supersamples comprises requiring all but one of the total number of supersamples to match.

19. The computer-readable medium of claim 17 wherein requiring the number of matching supersamples comprises requiring all but two of the total number of supersamples to match.

20. The computer-readable medium of claim 17 wherein requiring the number of matching supersamples comprises requiring all supersamples to match.

21. A computer-readable medium embodying machine instructions implementing a method for determining groups of near-duplicate items in a search engine query result, comprising, for each of two items being compared:

combining four samples of features into each of six supersamples;

recording each supersample to 16 bits of precision; and

requiring four of the six supersamples to match.

22. A computer-readable medium embodying machine instructions implementing a method for determining groups of near-duplicate items in a search engine query result, comprising, for each of two items being compared:

combining four samples of features into each of seven supersamples;

recording each supersample to 16 bits of precision; and

requiring five of the seven supersamples to match.